



Date : 18/06/2008

21 世纪的标题表: lcsh-es.org 双语数据库

Michael Kreyche

系统图书馆员, 肯特州立大学

美国, 俄亥俄州, 肯特市

中文翻译: 喻菲 (中国国家图书馆)

Chinese translator: Yu Fei,

(National Library of China)

Meeting:

129. Classification and Indexing

Simultaneous Interpretation:

English, Arabic, Chinese, French, German, Russian and Spanish

WORLD LIBRARY AND INFORMATION CONGRESS: 74TH IFLA GENERAL CONFERENCE AND COUNCIL

10-14 August 2008, Québec, Canada

<http://www.ifla.org/IV/ifla74/index.htm>

摘要

西班牙语是世界上使用最广的语言之一, 对西班牙语标题表研究的回顾, 显示了人们长期以来在这个领域所做的大量努力。这些努力通常是通过合作形式开展的, 但很大程度上还是彼此独立的。如果不考虑其它障碍, 技术进步使我们有可能进行更大程度的合作, 有益于国际图书馆界的发展。lcsh-es.org 项目用实践的方法论证了这个观念, 提出了一个规范控制的国际合作新模型。项目网站访问地址: <http://lcsh-es.org>。

西班牙语和标题表

众所周知, 西班牙语是世界上使用和地理分布最广的语言之一。估计一下说西班牙语的人数, 来自大不列颠百科全书数据有 3.58 亿¹人口说西班牙语; 来自 *Ethnologue*² 的数据有 27 个国家 3.22 亿人口说西班牙语。这里的西班牙语指的是卡斯蒂利亚语 (Castilian), 它起源于来自伊比利亚半岛的几种地方方言和相关语言中的一种。

如今，讲西班牙语的大量人口居住在美洲，分布在南美洲的大部分地区，加勒比海的部分地区、中美洲、墨西哥和北美洲。在过去的 20 多年时间里，西班牙语人口大量移民入美国；很多大型国家连锁商店里同时设立英语和西班牙语的标牌成了一种规范作法。因此，在美国的图书馆里为讲西班牙语的用户服务变得十分重要就毫不为奇了。很多美国图书馆在他们的网站上设有西班牙语的页面和西班牙语的联机目录用户界面，但是提供包含西班牙语书目记录或者是带有西班牙语标题表的目录就没那么常见了。不幸的是，我们还没有发现哪个单一的标题资源能像国会图书馆标题表那样内容全面、技术领先或易于使用。

在 OCLC 发现的最早出版的西班牙语标题表，来源于最初在 1986 年为公共图书馆出版的标题列表³。这个标题列表的第 4 版印刷版 1996 年由高等科学研究委员会（Consejo Superior de Investigaciones Científicas，CSIC）出版⁴，第 5 版以光盘形式于 2000 年出版。很显然，西班牙国家图书馆（Biblioteca Nacional de España，BNE）的规范文档最初出版于 1996 年⁵，也是以光盘形式出版⁶。近年以来，西班牙国家图书馆和高等科学研究委员会都在网上公开提供他们的规范文档。

在最近出版 2004 年美国图书馆协会年会双语卷上，包括了一个项目的论文以及两篇特邀论文，可以方便获得关于美国开发的西班牙语标题表的相关信息。⁷其中 4 篇论文论述了基于国会图书馆标题表编制的 3 种最有影响力的标题表中的一种或几种。

其中的一个标题表是《图书馆资料标题表》（Lista de Encabezamientos de Materias para Bibliotecas，LEMB），最初由泛美联盟（现在叫美洲国家组织）赞助出版于 1967 年。该标题表现在由哥伦比亚共和国银行的阿朗哥图书馆（Biblioteca Luis Ángel Arango of the Banco de la República in Colombia）维护，可通过购买获得其电子格式的数据。

《资料标题表》（Lista de Encabezamientos de Materias，LEM）最初也是出版于 1967 年。《资料标题表》一般被称为《Escamilla 标题表》（Escamilla list），以它的编制者墨西哥国立图书馆的 Gloria Escamilla 的名字命名。尽管目前该表仍由图书馆维护，但迄今为止出版的修订版本仅为 1978 年出版的第二版。在布依诺斯艾利斯举办的第 70 届 IFLA 大会上，来自墨西哥国立自治大学（Universidad Nacional Autónoma de México）的 Filiberto Felipe Martínez Arellano⁸提议开发这个标题表的新版本。

第 3 个标题表是 Bilindex 标题表，1986 年作为公共基金项目出版。Floriscanto 出版社的商业投资运作使其得以出版至今。Bilindex 的最初版本基于 LEMB 标题表、LEM 标题表和许多其它的资源。值得一提的是，LEMB 标题表的最初编制者看到过 LEM 标题表的早期出版前版本。这种开放共享在随后的几年出现了问题，部分原因是 LEMS 标题表和 Bilindex 标题表的商业化和人们对保护知识产权的关注。另一个原因可能是，在 19 世纪 80、90 年代采用的维护规范数据的自动化系统的封闭性特点。

最近在墨西哥有两个值得关注的合作项目。一个项目由墨西哥学院（Colegio de México）

主持, Martínez Arellano 在 SALSA 报告中曾经提到。这个项目非常有名, 很多资料都提到过它。⁹另外一个项目相对较新, 这个项目的协调人, 墨西哥圣路易斯波托西自治大学 (Universidad Autónoma de San Luís Potosí) 的 Julia Margarita Martínez Saldaña 描述了这个项目的具体细节。这个合作项目的的主要内容就是参加国会图书馆的“名称规范和主题规范合作项目”(NACO 和 SACO¹⁰), 而不是西班牙语标题表。项目开发了一个可以在 3 个不同的系统之间交换规范记录的界面, 可以克服封闭性自动化系统的一些局限。

2004 年, LFLA 大会在布依诺斯艾利斯召开。这次会议激发了拉丁美洲国家对《资源描述和检索》(Resource Description and Access) (该编目规则预计在 2009 年采纳) 的兴趣, 一系列针对该规则的编目人员年会也随之召开。此外, 为了鼓励广泛的讨论新的编目规则, 这些会议还讨论了有关编目的其它议题, 如规范控制等内容。2005 年, 在秘鲁首都利马召开了第一届国际编目会议 (The first Encuentro Internacional de Catalogadores)。会议内容包括几个报告会、一个小组讨论会和一个关于规范控制的讨论会。¹¹会议论文已印刷出版。¹²第二届国际编目会议 2006 年在墨西哥城召开。就是在这次会议上首次宣布 lcsch-es.org 项目。这次会议的论文同样也已出版。¹³第三届国际编目会议 2007 年 11 月在布依诺斯艾利斯召开。会议征集了 8 篇关于名称控制和主题控制的论文, 这些论文都发布在该会议的网站上。¹⁴

在美国, 旧金山公共图书馆 (San Francisco Public Library) 因在目录中使用了西班牙语标题表为其它编目人员提供参考, 而长期倍受关注。最近, 纽约皇后区公共图书馆 (Queens Borough Public Library in New York) 承担了一个项目, 创建它自己的西班牙语标题表。这个项目的大部分工作都建立之前的旧金山项目的基础上。

得益于上述两个图书馆的合作, 以及同西班牙国家图书馆(BNE)和西班牙高等科学研究委员会(CSIC)的友好关系, lcsch-es.org 项目正在建立双语标题表数据库, 以实现西班牙语主题规范工作更广泛的国际合作。实现这个目标面临着诸如技术的、语言的、地理的、政治的、经济的种种困难, 而这其中最容易克服的困难可能就是技术困难了, 所以我们放在首位解决。我们希望专门设计的现代化的工具可以帮助克服其它的困难。

技术环境

机器可读和可共享的规范数据已经存在了大约 25 年了。然而, 在极大程度上, 规范文档存在于单个图书馆或一批图书馆目录的环境中, 很难在这种环境之外共享工作。此外, 图书馆目录一般来说是封闭的, 在这种系统中, 专门的系统被设计完成独有的任务。这种封闭的环境正在迅速的过时。

在过去的十年里, 开源软件团体开发了一种可在不同的硬件平台上运行的、令人称奇的、高质量、常规用途软件工具。同时, 技术专家也在努力开发数据格式的开放标准和通信协议。这带来了常被不严格的称为“Web 2.0”的创造性尝试的繁荣。

现在，因受到 Koha 系统¹⁵和 Evergreen 系统¹⁶成功的影响，以及 2007 年 CODE4LIB 大会（2007 CODE4LIB conference）的影响，图书馆界正在致力于“下一代编目”的研究。CODE4LIB 大会针对 Lucene 和 SOLR 软件工具¹⁷和 Casey Durfee 主持的项目“Open-Source Endeca in 250 Lines or Less”¹⁸（Endeca 是一个用来建立新型目录的商品）在大会前召集了一个专题讨论会。

尽管这些系统带有图书馆老主顾的设计观念，但仍是一个应用这些技术来满足图书馆馆员特殊需求的大好机会，这一点可在 lcsh-es.org 项目中得到证明。另一个实例就是，德国不伦瑞克卡洛罗-威廉娜理工大学（Technische Universität Carolo-Wilhelmina zu Braunschweig）的 Bernhard Eversberg 建立的国会标题表数据库。¹⁹这真是有些讽刺，当一些图书馆员正在怀疑维护结构化词表的价值的时候，结构化词表却引起了语义网研究者的关注。事实上，图书馆员的怀疑是有一些道理的，让我们更加有理由去反思规范工作和依靠技术用更有效的方式来进行规范工作。

lcsh-es.org 项目

lcsh-es.org 项目是建立在这样一个前提下：规范工作的国际合作的新时代的环境业已成熟，带来这个新时代的新型工具也都开发出来（特别是对于西班牙语标题表）。一段时期以来，这种新型工具的特征开始变得明显。从 2005 年 9 月到 2007 年 9 月，有三个特征在开发的最初阶段的目标中显露出来。

广泛性：在基于国会图书馆标题表的主题词表系统的不同语言中，西班牙语可能是很独特的，因为它地理分布很广，带有许多国家性和地区性的差异，也因而有五个而不是一两个构建好的带有语言变异和语言混合的主题词表系统（BNE, CSIC, LEM, LEMB 和 Bilindx）。不考虑目前这种分化是否令人满意，在西班牙语规范工作中需要参考这些资源这一点是十分明确的。所以从多种资源中收集同样多的标题是一个很重要的目标。

高效和便于使用：广泛性必然包含着效率，但这项特征自身就意味着使其他人的工作更简单、更高效。来自互联网的一个经验就是，网络工具因为简单和便于使用而在它们出现的早期便得到人们的使用，而也许在这个时期这些工具在其它方面还存在缺陷。在 lcsh-es.org 系统中，检索十分方便，结果显示也十分迅速。同时进行了西班牙语词汇和英语词汇的标引，也可以进行基本关键词检索。

免费在网上使用：这个特征和高效和便于使用特征十分相关。要得到广泛的接受，唯一的要求就是需要像其它的网络服务一样的浏览程序，特别是启动程序；可以免费使用也是人们的预期。

基于 3 个测量指标的成功，lcsh-es.org 系统通过了第一阶段的验证。第一个测量指标是集成资源的数量。原始数据来自旧金山公共图书馆的书目记录，该图书馆从参加最初的 Bilindex 项目开始就在目录系统中使用了西班牙语标题表而长期享有声誉。多年以来，其它图书馆的编

目人员也在参考它的目录来查找西班牙语标题。将西班牙语标题对应到相应的英语标题就构成了一部词典，可以更加简单的查询西班牙语标题在国会图书馆标题表中的对应标题。第二个数据来源是纽约皇后区公共图书馆，他们正在建立他们自己的西班牙语标题表。还有一些标题从 CSIC 光盘和其它从 BNE 下载的目录中抽取。最后，从最初的 Bilindex 的英语—西班牙语索引中抽出的词使 lcs-h-es.org 系统最终的记录数量达到 50,000 条。另外一个有用的数据资源是下载的国会图书馆标题表文件的 Fred 2.0²⁰副本，该副本由 Simon Spero 公布（这个数据源同样在上面介绍的项目中被 Eversberg 使用）。尽管它不能为数据库提供新的西班牙语词汇，但是它可以验证在其它资源中发现的国会图书馆标题表中的词汇。

第二个测量指标的成功是稳定增长的用户数量。来自网络服务器的访问日志，2007 年至少访问了 100 个页面的用户中，所有的排名前 40 的机构或网络中，排名前 12 名的用户访问量都超过了 1000 个页面。这些排名前 40 名的用户一共访问了大约 35,000 个页面，其中墨西哥和西班牙（主要是学术图书馆）各 10,000 个页面，其余的是美国（主要是公共图书馆）15,000 个页面。即使现在的数据库还远未达到广泛性的要求，但是它达到了一个关键性的规模，这使得它能在多种类型的图书馆中使用。来自西班牙的页面访问有些让人吃惊，因为所有可确认的页面访问的机构都在加泰罗尼亚地区，而在这个地区西班牙语并不是第一使用语言。一个可能的解释就是那里的图书馆员使用数据库查询国会图书馆标题表的西班牙语的词汇，然后将它翻译成加泰罗尼亚语。如果是这样的话，这表示有可能建立一个三种语言的数据库！

第三个测量指标的成功是初始阶段的工作帮助我们获得了国家人文基金会（National Endowment for the Humanities, NEH）对目前正在开发阶段（2007 年 9 月—2008 年 9 月）的资助。大部分资助的基金都用来雇佣学生程序员，加快开发的速度。这个阶段的主要目标是使数据库具有另外两个特征。

基于 MARC 格式：数据被转换成 MARC21 格式，数据结构相应进行了修改。这些转换将会使输出结果立即有用，有助于把来自其它资源的数据整合成 MARC。迄今为止（2008 年 4 月），NEH 资助的大部分工作都在致力于这方面的内容。

合作性：数据库必须拥有一个合作和交互的基础平台，我们已经开发了一个登录系统来支持这项功能。数据库的用户可以通过不同的方式为数据库建设作出贡献：找出错误、建议修改、添加 4xx 字段的词汇或者 670 字段和 675 字段的内容。数据库系统应该可以通过修改一条已存在的记录的 1xx 字段和 4xx 字段的词汇来创建一条新的规范记录，以反应地区或国家的用法。这个系统的目标是建设一个编目人员的社区，为每个人的规范工作提供帮助。一旦拥有了足够多的参与者，这样一个网络系统就可以成为一个使用标签技术（techniques of tagging）和分众分类法（folksonomies）构建结构化词表的案例。

同时，一些其它的特征使数据库更便于使用和更加实用。例如，“见”（又或者是“参见”）关系现在整合进了数据库；简单点击鼠标就能拷贝和复制主题标目并能高亮显示（不用再从头

至尾拖动光标)。目前正在计划在词典和其它互联网参考工具之间添加链接, 以及通过一个简单的输出功能可以下载经过测试的单条 MARC 记录。将来也可能添加一个功能, 将记录直接发送到本地系统, 同样的编目人员也可以从 OCLC 直接下载记录。

许多个性化的特征也在考虑之中。其中一个个性化特征就是编目人员可以指定个性化检索结果的首选标题资源。另外一个特征是为复制和粘贴设定符合本地系统的字符编码和 MARC 内容标识。

除了改善手动工作的用户界面, 批量处理和 Web 服务的能力可能会十分有用。例如, 一个没有西班牙语标题的图书馆可以上载一个书目记录文件, 然后安排一个批量处理程序在数据库中检索英语标题并添加等同的西班牙语词汇。一个 Web 服务程序可以按逐条标题的形式执行同样的功能, 自动处理新记录输入系统的任务。这需要在本地图书馆系统中进行常规设计, 或许图书馆使用开放资源目录可以鼓励在更大范围中开发这种功能。这种服务也许对一些非图书馆网站也同样有用。

结束语

展望未来, 希望 lcs-h-es.org 项目促进的非正式合作将发展成为建立在不断发展的基础上数据共建的正式承诺和协议。为了集合所有的主要西班牙语标题表, 几乎是肯定的需要一些充足的创始基金以满足商业生产者的需要。同一些主要图书馆合作能够有效的吸引这样的资金, 并不是不切实际的。开发一个持久的经济支持模型将是更大的挑战。

即使这种做法被证明难以实现, 这个项目也可以激励图书馆员采用相似的技术或者同样的代码库, 用其它语言工作。事实上, 更切实际的做法是开始开发一个尚不存在或者即将存在的基于国会图书馆标题表的某种语言的词表。这个系统采用 Apache 网站服务器, MySQL 数据库和 PHP 脚本语言, 这些都可以在许多的操作系统上运行。原始数据库采用拉丁-1 字符集, 但是现在的版本存储的所有记录都采用的 UTF-8 字符集, 因此事实上这个系统是有可能采用任何文字和任何语言的。

最后, 一些基本问题也值得思考。第一, 是国会图书馆标题表自身的缺陷问题。国会图书馆标题表长期以来被人们认为存在文化的和语言的偏见, 它的不一致性也是很多问题产生的根源。全新地、系统地看待这些问题, 可以产生创建一个国际化或通用的西班牙标题体系许多好的想法。也许可以开发一种双语国会图书馆标题表。该表分成两个部分, 其中的一个开发成新的国际化版本作为原始版本的参照。原始版本只要有用的可以继续以保守方式开发, 即使将来这个版本变得僵化, 或者停止使用。

另外一个问题就是建立规范文档本身是否有意义。当目录印刷在卡片、书籍甚至是缩微胶片上时, 一个确定的概念使用单个的规范词是有用的。而现在存储器是最便宜的计算机资源, 也许规范记录的 1xx 字段应该消失, 其内容作为 4xx 字段的标签, 有效的建立一组符合地理用

法或知识领域的同义词。没有任何技术上的原因会影响到在合适的书目记录中包括同一概念的多个词汇。同样也没有功能上的原因，因为针对同义词中任何一个词的检索，都会检索到所有的书目记录。只要我们图书馆的资助人有兴趣，他们总能选择出“合适”的词汇。

无论如何，比较过去和现在，主题词表的未来将产生巨大的变化。当书籍通常都以电子格式出版，我们现在所熟知的图书馆里的主题分析将不复存在。随着电子文档数量的增长，我们希望看到也许有更复杂的计算机软件来完成这些工作，图书馆员或主题分析专家仅需要提供一点点辅助工作。全文检索毫无疑问是一种高效的技术，但是人们倾向于开发强有力的分类技术，因此结构化词表可能以某种方式继续存在。

¹ "Spanish language." Encyclopædia Britannica. 2008. Encyclopædia Britannica Online. 7 Apr. 2008 <http://search.eb.com/eb/article-9068992> (accessed April 9, 2008).

² Gordon, Raymond G., Jr. (ed.), 2005. *Ethnologue: Languages of the World*, Fifteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/show_language.asp?code=spa (accessed April 9, 2008).

³ Lista de encabezamientos de materia para las bibliotecas públicas. Madrid: Ministerio de Cultura, Dirección General del Libro y Bibliotecas, 1986.

⁴ Unidad de Coordinación de Bibliotecas (C.BIC). Lista de Encabezamientos de materia de la Red de Bibliotecas del CSIC.,4a. ed. Madrid: Consejo Superior de Investigaciones Científicas, 1995.

⁵ "Autoridades de la Biblioteca Nacional de España." *Correo Bibliotecario*. Valladolid: Biblioteca de Castilla y León. Sumario 12, enero-febrero 1997. http://www.bcl.jcyl.es/CORREO/plantilla_neditorial.php?id_neditorial=63&id_seccion=5&RsCorreoNum=12 (accessed April 9, 2008).

⁶ Autoridades de la Biblioteca Nacional. Madrid: Chadwyck-Healey España, 1996.

⁷ Miller, David and Filiberto Felipe Martínez Arellano, editors. *Salsa de tópicos = Subjects in SALSA: Spanish and Latin American subject access*. Chicago: Association for Library Collections & Technical Services, American Library Association, 2007. ALCTS papers on library technical services & collections; # 14.

⁸ Martínez Arellano, Filiberto Felipe. *Desarrollo de una lista de encabezamientos de materias en Español*. Buenos Aires: World Library and Information Congress, 70th IFLA General Conference and Council, 22-27 August 2004. http://www.ifla.org/IV/ifla70/papers/039s_trans-Arellano.pdf (English: <http://www.ifla.org/IV/ifla70/papers/039e-Arellano.pdf>; French: http://www.ifla.org/IV/ifla70/papers/039f_trans-Arellano.pdf; Russian: http://www.ifla.org/IV/ifla70/papers/039r_trans-Arellano.pdf, (accessed April 9, 2008).

⁹ Quijano-Solís, Alvaro, Pilar María Moreno-Jiménex, Reynaldo Figueroa-Servín. "Automated Authority Files of Spanish-Language Subject Headings." *The LCSH Century: One Hundred Years with the Library of Congress Subject Headings System*. Haworth Press, 2000, p. 209-223.

¹⁰ See also the following paper from IFLA70: Cristán, Ana. *The SACO Program in Latin America*. Buenos Aires: World Library and Information Congress, 70th IFLA General Conference and Council, 22-27 August 2004. <http://www.ifla.org/IV/ifla70/papers/040e-Cristan.pdf>; German: http://www.ifla.org/IV/ifla70/papers/040g_trans-Cristan.pdf; Russian:

http://www.ifla.org/IV/ifla70/papers/040r_trans-Cristan.pdf; Spanish:

http://www.ifla.org/IV/ifla70/papers/040s_trans-Cristan.pdf (accessed April 11, 2008).

¹¹ Conference program, <http://bvirtual.bnp.gob.pe/inscripcion/programacion.htm> (accessed April 12, 2008).

¹² Centro Bibliográfico Nacional . Nuevas tendencias en la normalización y sistematización de la información : ponencias y conclusiones. Lima: Biblioteca Nacional del Perú, Fondo Editorial, 2006.

¹³ Martínez Arellano, Filiberto Felipe and Ariel Alejandro Rodríguez García, compilers. Memoria del Segundo Encuentro Internacional de Catalogación : tendencias en la teoría y práctica de la catalogación bibliográfica, 12 al 14 de septiembre de 2006. México : UNAM, Centro Universitario de Investigaciones Bibliotecológicas, Instituto de Investigaciones Bibliográficas; Library Outsourcing Service, 2007

¹⁴ III Encuentro de Catalogadores. http://www.bn.gov.ar/ACT_EjesEncuentro.aspx (accessed April 12, 2008).

¹⁵ <http://www.koha.org/> (accessed April 12, 2008).

¹⁶ <http://open-ils.org/> (accessed April 12, 2008).

¹⁷ <http://code4lib.org/node/139> (accessed April 12, 2008).

¹⁸ <http://code4lib.org/2007/durfee> (accessed April 12, 2008).

¹⁹ <http://www.biblio.tu-bs.de/db/lcsh> (accessed April 12, 2008).

²⁰ Spero, Simon. Fred 2.0: Cosmos, Taxis, and the Future of Bibliographic Control. <http://www.ibiblio.org/fred2.0/wordpress/> (accessed April 12, 2008).